DOCUMENT RESUME

ED 243 306 FL 014 310

AUTHOR Perkins, Kyle; Parish, Charles
TITLE Direct Versus Indirect Measures of Writing Proficiency: Research in ESL Composition.
PUB DATE Apr 84
NOTE 35p.; Paper presented at the Annual Meeting of the Illinois Teachers of English to Speakers of Other Languages/Bilingual Education (Chicago, IL, April 6-7, 1984).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *English (Second Language); Higher Education; Language Processing; *Revision (Written Composition); *Second Language Instruction; *Sentence Structure; *Writing Evaluation; *Writing Skills

ABSTRACT

A comparison of measures of the attained writing proficiency of 45 college-level students of English as a second language is reported. Students were tested by two indirect measures, the Test of the Ability to Subordinate (TAS) and the Revision and Editing Test (RET), and their compositions were evaluated by a direct measure, holistic evaluation. The TAS consists of 45 questions requiring students to combine two or more sentences with any form of subordination. The RET asks questions requiring an understanding of units larger than the sentence and the ability to improve sentences or sentence groups. Results of the two tests were compared with the holistic composition evaluation, the criterion measure, and it was determined that: (1) the TAS has higher reliability and validity than the RET for these subjects; (2) neither test has concurrent validity with the holistic evaluation; (3) the revision and editing subparts are of equal difficulty for this group but assess different constructs as they were designed to; and (4) this study's findings on the holistic evaluation's composition profile differ markedly from previous researchers' conclusions. Despite finding number 2, it was concluded that the holistic scoring profile guide and RET provide similar criterion-related information for assessment, placement, exemption, certification, and promotion. Statistical analysis results and references are included. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED243306

DIRECT VERSUS INDIRECT MEASURES OF WRITING PROFICIENCY: RESEARCH IN ESL COMPOSITION

Kyle Perkins and Charles Parish

Department of Linguistics
Southern Illinois University
Carbondale, Illinois 62901

Office: (618) 536-3385 (618) 536-3385

Home: (618) 985-6056 (618) 549-1398

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Kyle Perkins
Charles Parish

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

FL014310

ABSTRACT

One of the currently researched issues in ESL/EFL composition is the extent to which indirect and direct measures of composition tap the same construct. This paper reports the results of an investigation in which two indirect measures, The Test of the Ability to Subordinate (TAS) and the Revision and Editing Test (RET), and one direct measure, holistic evaluation of a writing work sample, were used to evaluate the attained writing proficiency of a group of ESL students enrolled in a university-level composition class. The TAS has 45 items, each of which contains two or three sentences which the student is asked to combine into one sentence by filling in missing words in a sentence frame. The test accepts as correct any standard type of subordination. The RET contains 15 questions on units larger than the sentence and 15 questions on how to improve sentences or groups of sentences. "The student must discover thesis and topic sentences; make judgments about level of language, voice, coherence, development, logic, diction; discern methods of development, errors of logic, unstated assumptions, sentence variety, patterns of errors, effectiveness of examples; and offer suggestions for revision at every level. Questions about sentences ask how best to combine or revise or punctuate or correct or reword with alternative choices of revised sentences to choose from."

The results indicate that for this particular subject pool and the measures employed in this study, neither of the indirect measures accounts for enough reliable variance in the criterion measure to replace it.

Introduction

There is considerable debate in the composition research community concerning the efficacy of direct and indirect measures of attained writing proficiency. Some researchers, like Michael and Shaeffer (1979) and Diederich (1974), advocate the use of indirect, multiple-choice tests of writing.

Michael and Shaeffer (p. 132) stated:

> A short multiple-choice examination requiring knowledge of principles about grammar, usage, choice of words and phrases, and syntax such as the TSWE (Test of Standard Written English) or a subset of the CSUC-EPT (California State University and Colleges English Placement Test) reflecting similar competencies provides a more accurate prediction of success in a basic English composition course emphasizing writing skills than does an essay examination [a direct measure] requiring the actual demonstration of writing skills.

In a similar vein, Diederich (pp. 79-80) noted:

> My final word of advice on such tests [objective tests] is not to despise them. Objective test items can easily, quickly and reliably test a students's knowledge of the rules and conventions of English. In writing, if a student is not sure that [s]/he knows how to use a certain construction [s]/he can

> change his/[her] sentence to avoid using it, but in an objective test, you can give him/[her] a sentence with that construction in it, and [s]/he has to decide whether it is correct or incorrect. Arguments over whether answering objective items 'is the same thing as' actual writing or speaking are futile. For that matter, writing one essay is not 'the same thing as' writing another essay, even on the same day; we have seen that even the most carefully determined grades on such essays rarely correlate higher than 0.70. Then the astonishing thing is that scores on a good objective test of English usage often correlate about 0.70 with averages of the two essay grades. It does not matter that they do not 'really' measure 'the same thing.' If students who are good at one tend to be good at the other, and vice versa, then it is a good indicator of proficiency in written English. Call it an editing test if you like, but I can promise you that students who do well on it tend to be good writers.

But other researchers like Braddock, Lloyd-Jones and Schoer (1963) and Breland and Gaynor (1979) have criticized objective tests for their lack of validity and the impressions about the low valuation of

writing which they convey.

Braddock, Lloyd-Jones and Schoer (p. 42):

> The most serious charge against multiple-choice tests is their lack of validity. Not only do they not require the examinee to perform the behavior being measured--[s]/he does no actual writing; but these tests also make no attempt to measure the 'large elements' of composition, even indirectly.

Breland and Gaynor (p. 127):

> Indirect measures lack face validity and credibility among members of the English profession and educators generally, and they tend to deliver a message to students that writing is not important.

While the arguments by Braddock, Lloyd-Jones and Schoer and Breland and Gaynor for direct measures of writing are convincing from the stand-point of validity, direct measures have been traditionally criticized for low reliability due to the graders' subjectivity. Jacobs et al (1981) list some of the more commonly cited threats to the reliability of direct assessment of writing: (1) readers may have different standards of severity (p. 24); (2) readers may distribute scores differently along the scoring scale; (3) readers may be inconsistent in applying the standards of the evaluation; (4) readers may react to certain elements in the evaluation or in the papers (p. 25); and (5) readers may value different aspects of a composition (p. 26).

(For further discussion on the correlation between direct and indirect test scores in composition, see Davis, Scriven and Thomas [1981]; Hogan and Mishler [1980]; Moss, Cole and Khampalikit [1982]; and Stiggins [1982].)

Research Questions

The purpose of this investigation is fourfold: the primary questions are to determine (1) the reliability and validity of two indirect and one direct measure of writing for an English as a Second Language (L2) subject pool, and (2) whether either of the two indirect measures can replace the direct measure. The secondary questions are (3) whether there is a difference in the performance on different tasks in one of the indirect measures (i.e. the RET) in order to determine component validity; and (4) which components of the rating criteria account for the most variance in the direct measures (i.e., the holistic scoring).

Procedures

Subjects. The subjects for this research were 45 undergraduate foreign students enrolled in Basic Composition for Foreign Students. Native languages represented in the sample included Arabic, Malay, Korean, Vietnamese, Thai, Chinese, Japanese, Greek, and Spanish. These students are placed in the course on the basis of their graduation from the intensive English program (non-credit) or by normal University placement procedures. The faculty for courses are trained in ESL procedures.

Materials and Procedures. During the last two weeks of a

fifteen-week semester, data were collected from the Test of Ability to Subordinate (TAS), the Revision and Editing Test (RET), and a composition test in which the subjects wrote an in-class essay which was graded holistically (HOLISTIC).

TAS. The TAS is considered, for this research, an indirect measure of writing, since it entails no text construction or actual composing. The TAS assesses an examinee's ability to write different kinds of English sentences. In each of the 45 items two or three sentences are given and the subject is asked to combine them into one sentence by filling in missing words in a sentence frame. For example,

a. They left the car. b. The car was parked outside.

Frame: They left the car _____.

Any standard type of subordination is acceptable as a correct answer, and no penalty is assessed for errors which do not interfere with the subordinating structure. The following description of the TAS is found in the Teacher's Manual (1978, p. 1):

> The Test of Ability to Subordinate was designed primarily as a diagnostic instrument to measure the ability of non-native speakers of English to control certain structures of subordination in writing. It was developed to fill the need for an easily scoreable, objective test that would measure students' ability to manipulate particular grammatical structures generally regarded as indicative of mature writing . . . the primary diagnostic purpose of the TAS is to identify which of nine structures of subordination students can manipulate and which they cannot, but

a careful analysis of responses will highlight difficulties, if any, with subject-verb agreement, plurals, tenses and articles. A teacher can use the information from the TAS to group students in terms of their strengths and weaknesses in subordination and to develop a writing program that will enable them to use these structures effectively as well as deal with other grammatical problems.

RET. The RET is an exit exam designed to assess a native English speaker's (L1) ability to revise and to edit. It was developed by Professor Carolyn B. Matalene and her freshman English staff at the University of South Carolina. Matalene (1982) describes the RET:

> The test is divided into two parts; the first fifteen questions are on units larger than the sentence, the last fifteen questions on how to improve sentences or groups of sentences. Thus revision precedes editing. The test is printed with the entire essay (sentences and paragraphs numbered) in reduced form on page one of the test booklet and stapled so that the student can easily flip back and forth to the essay. Answers are recorded on a computer form to be machine-graded. A perfect score is 150, or thirty times each five-point answer. Passing should theoretically be 120, or thirty times each four-point answer, since both the four- and the five-point answers are correct according to departmental standards. The questions ask students to discover

> thesis and topic sentences; to make judgments about level of language, voice, coherence, development, logic, diction; to discern methods of development, errors of logic, unstated assumptions, sentence variety, patterns of errors, effectiveness of examples; and to offer suggestions for revision at every level. Questions about sentences ask how best to combine or revise or punctuate or correct or reword with alternative choices of revised sentences to choose from. In short, students are asked how best to revise and edit the entire essay. Because of the test format, questions sometimes ask what error or problem is the most serious; in this regard the bias of the test is that structural errors are more important than surface ones, that editing is more important than proofreading (p. 373).

The writers examined the answer key--5 points for the best of five possible answers, with 1 point for the least appropriate choice--and agreed with the test developer's decisions. Matalene (1982; pp. 373-374) offered the following examples:

1. Comment on the writer's use of language in this essay;

   A. The writer uses language which is unusually formal for his/[her] topic. B. The writer's language is too informal for this topic. C. The writer's language is sometimes archaic and therefore inappropriate. D. The writer uses language which is often vague and general. E. The writer mixes conversational language with formal language.

2. In this essay, which of the following errors of logic is the most serious: A. The writer attributes too many possible results to one cause. B. The writer uses a false syllogism. C. The writer uses a false comparison. D. The writer's generalizations are not proven. E. The writer uses undefined terms.

The authors used Form C of the RET for this investigation.[1] The

---

[1]We are indebted to Professor Matalene for her kind permission to use one of the forms of her test for this research, since it is not yet available commercially.

---

decision to use a test designed for native speakers (The RET) as well as one used for foreign students of English (The TAS) was based on two points: 1) the RET, as an exit exam, is a criterion-referenced instrument, enabling us to assess the proficiency of our students according to an external standard, and 2) it provided us with a basis for measuring the achievement of the foreign student in a comparable course.

Holistic. The subjects were assigned a composition which was written in class under test conditions: the subjects were allowed 50 minutes to write the exam. Five topics allowing an expository-argumentative focus were offered in the following format:

1. The role of agriculture in my country today.
2. The most difficult problem facing my country today.
3. The effects of modern technology on the way of life in my country.

4. Why young people in my country need a college education.

5. Changes that I would like to see in my country.

Although these topics may seem to the American eye to be dull or banal, the actual outcome was that the great majority of the papers showed the writers to be involved and committed. To varying extents, all allow the writers to be informative about their culture and in some--2, 4, 5--even polemical. This provides a motivation to exploit the topic thoroughly.

The researchers graded the compositions using a holistic approach and a scoring guide. The researchers selected holistic evaluation of a direct writing sample as the criterion measure for this investigation because holistic scoring of a writing work sample--the researchers believe--has the highest construct validity of all of the composition evaluation schemes available today when overall attained writing proficiency is the construct to be assessed. Because holistic scoring evaluates a whole text, the grader(s) can determine (1) whether a thesis has been clearly stated, developed, and supported and whether an issue has been clearly raised and sufficiently resolved; (2) whether sufficient support and development (i.e., redundancy in the communications science sense) have been provided for the reader; (3) whether the writer has tried to accommodate the reader in terms of how much (or how little) the writer thinks the reader knows about the subject (i.e., has the writer tried to accommodate the intended audience?); (4) whether there is grammatical and lexical cohesion and overall coherence; (5) whether one paragraph is "hinged" smoothly to preceding and succeeding paragraphs through tne accomplished use of transitional words, phrases, and structure vocabulary; (6) whether new information has been converted into given information

that forms the base for predicating additional new information; and (7) whether the piece "makes sense."

The researchers used Jacobs et al. (1981) ESL Composition Profile as the scoring guide for the holistic evaluation. Jacobs et al. describe the Profile form:

> The PROFILE form . . . contains five component scales, each focussing on an important aspect of composition and weighted according to its approximate importance for written communication: content (30 points), organization (20 points), vocabulary (20 points), language use (25 points), and mechanics (5 points). The total weight for each component is further broken down into numerical ranges that correspond to four mastery levels: excellent to very good, good to average, fair to poor, and very poor.(p. 31)[2]

---

[2]This "rating schedule"--that is, both the point allocation for the 5 components and the criteria specified for the 4 mastery levels--may seem at first glance to constitute analytical scoring rather than holistic scoring. A closer look will show that this is really a sensible attempt to standardize holistic scoring. If the reader will remember that it has sometimes been--and certainly still is in many English departments--the practice to assign the grade of B to a paper with a comma splice, run-on sentence, or sentence fragment, and the grade of C (or lower) to a composition containing two of these faults,

and a failing grade to one containing three or more, it will be clear that criteria such as these, with limits assigned to the value of various components, are necessary to prevent the abuses of impressionistic holistic grading.

The criteria for the grading range of the components--for example, "Vocabulary"--are specified to help the grader make decisions:

20-18 EXCELLENT TO VERY GOOD: sophisticated range; effective word/idiom choice and usage; word form mastery; appropriate register

17-14 GOOD TO AVERAGE: adequate range; occasional errors of word/idiom form, choice, usage but meaning not obscured

13-10 FAIR TO POOR: limited range; frequent errors of word/idiom form, choice, usage; meaning confused or obscured

9-7 VERY POOR: essentially translation; little knowledge of English vocabulary, idioms, word form; OR not enough to evaluate.

The two researchers graded each composition independently and anonymously, interpreting and applying the Composition Profile as rigorously as possible. The grading was counterbalanced to preclude any halo effect and contaminating bias. Each paper received a holistic score from both researchers, and the two scores were averaged to produce the holistic score. Inter-rater reliability, as shown in Table 1, was .70.

Results and Discussion

(Table 1 about here)

Question 1: Reliability and Validity (of All Measures)

Of the three measures, the TAS exhibited the highest estimate of observed reliability, 0.78. The interrater reliability of the two investigators for the holistic evaluation was 0.70, and the RET had an observed reliability of 0.56. The difference in reliability between TAS and RET may have been due to the number of items in each test (TAS, 45; RET, 30), because reliability is directly affected by the number of items in a test. All other things being equal, a test with more items provides for a greater range of scores, and consequently subjects are more widely dispersed along the scoring continuum. With this in mind, the researchers used the Spearman-Brown Prophecy Formula to estimate the reliabilities of the TAS and RET, if they were lengthened by the addition of similar items to a total of 100 items. The estimates of 100-item reliability, 0.89 and 0.81, approach the reliability coefficients of commercially-produced standardized tests. The Spearman-Brown Prophecy Formula indicated that 2.43 raters would be needed to achieve an interrater reliability of 0.85 for the holistic evaluation.

The lower estimate of reliability for the RET may also be explained by the fact that it was designed for an L1 mainstream population. The L2 subject pool employed for this investigation still struggle with writing grammatically correct and complete sentences, with synthesizing the notions of canonical organization of introduction, body and conclusion, and with the notions of thesis statement, topic sentences, and supporting details. In other words, the RET assesses a level of

sophistication in the written code which this subject pool are not yet prepared for. The point being made here in testing jargon is that the RET may have been too difficult for this subject pool, and test difficulty can affect a test's reliability. When a test is too difficult, scores are compacted at one end of the scoring continuum, resulting in what is called boundary effects. When scores are compacted, one cannot distinguish between persons, and there is, as a result, a loss of person separability or reliability.

(Table 2 about here)

Three different methods were employed to obtain validity coefficients for the two indirect measures: maximum validity possible, internal construct validity proportion, and correlation coefficient with the criterion measure. Henning (in press) gives the various derivations for the maximum validity possible coefficient, but for purposes of this paper "the maximum validity possible in the most ideal situation is equal to the square root of the reliability of the test, or $\sqrt{r_{tt(x)}}$."

The internal construct validity proportion method follows standard procedures--quite complex--and is also discussed in Henning. The results are shown also in Table 2.[3] For the third validity coefficient

---

[3]This procedure assumes that if the TAS items have construct validity, the point-biserial correlation between each TAS item and the total scores for TAS should be higher than the point-biserial correlations of the same items with the total scores for the holistic

evaluation, mathematically expressed as $r_{pbi_{TAS_i}\ TAST} > r_{pbi_{TAS_i}\ HOLT}$, where TAS_i refers to any item on the TAS, TAST refers to TAS total score, and HOLT refers to holistic total. The symbols indicate that the point-biserial correlations of individual TAS items with TAS totals should be greater than the point-biserial correlations of the same items with holistic totals. The point-biserial correlations between items and their own totals were corrected for part-whole overlap, because test items exhibit spuriously high correlation with their own totals since they make up a portion of that total. The single coefficient under the internal construct validity proportion column represents the proportion of items that correlated higher with their own totals (after part-whole overlap correction) than with holistic totals. For the RET, four- and five-point answers were coded as 1, because such answers are correct according to Matalene's departmental standards, and three-, two-, and one-point answers were coded as 0. The mean point biserial correlations were the following:

$\bar{X}\ TAS_i\ TAST = .27$ (corrected for part-whole overlap)

$\bar{X}\ TAS_i\ HOLT = .21$

$\bar{X}\ RET_i\ RETT = .02$ (corrected for part-whole overlap)

$\bar{X}\ RET_i\ HOLT = .32$

(Col. 3), the TAS has higher item discriminability than the RET and also tends to correlate higher with the criterion.

By all three methods of estimating validity, the TAS exhibits more validity than the RET with this population. This finding is

probably due to the fact that the TAS was designed for an L2 population and the RET was not. In addition the difficulty of the RET probably reduces its validity, because, as stated previously, reliability and validity are directly related. It would have been surprising to find that either test had an internal construct validity proportion of 1.00. Such a finding would indicate that the RET and TAS do not share variance with the holistic scores--which is not the case.

(Tables 3, 4, 5, 6 about here)

Question 2.

To determine whether one or both of the indirect measures (TAS or RET) could replace the direct measure (the holistic evaluation), the researchers employed stepwise multiple regression and Pearson product-moment correlation analyses. The TAS accounted for the most variance in the holistic scores and in grader one's holistic scores. The largest multiple R was between the TAS and grader one's holistic scores, 0.25. For grader two's holistic scores, the independent variables entered the regression equation in reverse order, i.e., with RET accounting for more variance than did TAS. The fact that the two predictor variables enter the equations in reverse order for the two different graders' holistic scores indicates that the two graders valued different aspects of the compositions.

Although all the F ratios are significant, the researchers could not decide on the basis of the multiple regression analyses whether one or both of the indirect measures could replace the direct measure. In all cases significant amounts of variance were accounted for, but after a thorough review of testing literature we could not isolate any

criterion for determining how large the R^2 statistic must be in order to conclude that two instruments are measuring the same constructs. Some papers enthusiastically endorse a multiple R of 0.36 while other researchers discount a multiple R of 0.70 as not meaningful.

It was decided to investigate the amount of shared reliable variance between the indirect and direct measures as an indicator of whether there was evidence of concurrent validity. Two different pieces of information are needed for this kind of analysis: reliability estimates and Pearson product-moment correlations of the measures.

(Table 6 about here)

Table 6 shows that the TAS and Holistic variables correlate at 0.44 and the RET and Holistic variables at 0.36, both significant at the .05 level for a two-tailed test. However, for research of this nature one must ignore the level of significance of the correlation coefficient and pay attention to the magnitude of the correlation for the following reason: with large sample sizes such as this one (df=43) it is possible for comparatively small correlation coefficients to be found significant; in fact, a correlation coefficient of 0.26 is significant at the .05 level for a two-tailed test with 43 degrees of freedom. As a result, coefficients of 0.50 or above are generally thought to be meaningful for research purposes.

And now to the issue of reliable shared variance between the indirect and direct measures of writing. Table 1 shows that the TAS had an observed reliability of 0.78; the Holistic evaluation, 0.70. The average reliability then for the two measures is 0.74. Table 7 indicates a correlation coefficient between the two variables of 0.44.

By squaring the correlation, we can say that the TAS accounts for 19 percent of the variability in the Holistic scores, $(.44)^2$=.19x100=19%. Squaring a correlation coefficient gives r^2, or the coefficient of determination. Since the two measures together have an average reliability of 0.74, the TAS accounts for 14 percent of the reliable variability in the Holistic scores: .74 x .19 = .14 x 100 = 14%.

The RET has an observed reliability of 0.56. The average reliability of the RET and Holistic variables is 0.63. The two tests correlated at .36. Squaring the correlation gives an r^2 (coefficient of determination) of .13. So the RET accounts for 13 percent of the variability in the Holistic scores. Since the two measures together have an average reliability of 0.63, the RET accounts for 8 percent of the reliable variance in the Holistic measure, .63x.13=.08x100=8%.

Conclusion:

Neither the TAS nor the RET accounts for enough reliable variance in the Holistic scores to replace it. It is true that method effect can account for as much as 50 percent of the common variance between two tests, but neither indirect measure comes anywhere close to accounting for the remaining 50% (or more) of the potential shared reliable variance.

Question 3:

The third research question concerns the subjects' performance on the two parts of the RET. The first fifteen questions are on units larger than the sentence ("revision"), the last fifteen questions on how to improve sentences or groups of sentences ("editing"). The

average score on items 1-15 was 53.22 with a standard deviation of 7.18; the average score for items 16-30 was 50.46 with a standard deviation of 6.02. An F ratio of 1.43, df 44,44, indicated that the variances were equal, and a t-value of 1.68, df 44 indicated that the means were not significantly different, i.e., the two halves of the test were of similar difficulty. As Table 6 indicates, the Revision and Editing sections of the RET correlated at -.06, so the two halves of the test clearly seem to be measuring different constructs. This factor is as consistent with an L2 population as with native speakers of English.

A further inspection of Table 6 shows meaningful correlations (above 0.50) between TAS and PROFILE language use which is not surprising, since both tap syntactic manipulation. Other meaningful correlations are shown for the RET subtests and RET total and PROFILE subparts and PROFILE total. Again, although these correlations are artificially high because the subparts contribute to the total score, we can at least conclude that the data demonstrate a considerable magnitude of concurrent validity.

(Table 7 about here)

Question 4.

The final research question addresses the amount of variance which each component scale (i.e., content, organization, vocabulary, language use, and mechanics) accounts for in the total Holistic score. Table 7 indicates that vocabulary, organization, and syntax together account for 94 percent of the variance in the total score. It is clear that these three components were valued by both raters with

consistency. What is surprising is the almost nil contribution to the accounting for variance by content, yet it is worth 30 points. There seems to be an almost total mismatch between the weight given the components by Jacobs et al. according to their approximate importance for written communication, and the variance which these components account for in the total, as the following set of ranks indicate:

	ENTERED REGRESSION EQUATION	JACOBS ET AL.'s WEIGHTING ACCORDING TO IMPORTANCE	
Vocabulary	1	3.5	20 pts.
Organization	2	3.5	20 pts.
Language Use	3	2	25 pts.
Content	4.5	1	30 pts.
Mechanics	4.5	5	5 pts.

The implication is that the two researchers and Jacobs et al. value aspects of a composition differently and react differently to certain elements in the evaluation process.

In summary, the answers to the four research questions are these: 1) the TAS has higher reliability and validity coefficients than the RET for this particular L2 subject pool. This finding is probably explained by the fact of the RET's level of difficulty as well as by the fact that it was designed to measure the attained proficiency of L1 students who had studied in an average freshman composition sequence. 2) Neither test accounts for enough reliable variance in the holistic evaluation to establish concurrent validity. It is probably unfair to consider

the TAS as a replacement for a proficiency measure because it was designed as a diagnostic measure of syntactic fluency for placement and diagnosis; the RET, however, is an exit exam and thus qualifies as a candidate for a proficiency test. 3) The Revision and Editing subparts are of equal difficulty for this subject pool but assess different constructs as they were intended to do. And finally, 4) there is a lack of concordance between the weights assigned to the PROFILE components and the order in which they entered the regression equation; i.e., the researchers' findings regarding which components account for most variance differ markedly from the decision of Jacobs et al.

Philosophical Reservations about the Measurement of Writing Proficiency.

Although the researchers set themselves the task of measuring the writing proficiency of this population under rigorous (but usual) test conditions, it is important to acknowledge the bias of this situation in relation to "real world" capability. There are at least three factors that should be remembered in order to maintain pedagogical and analytical perspective and fairness. First, only one composition was elicited for the holistic evaluation. Multiple writing samples might have been elicited from the subject pool, preferably at various times on different topics from different modes of discourse. Previous research has shown that attained proficiency in writing varies with topic and time of day.

Second, in having the subjects write the compositions, the researchers did not allow the subjects sufficient opportunity to follow all the steps prescribed in most models of the composing process. Most models have at least four stages: prewriting, writing, rewriting,

and editing. If such a model with these four stages had been followed with the subjects' instructors or the researchers providing sufficient guided revision, the subjects could have transformed their early drafts, which can be labelled Writer-Based prose, to more refined subsequent drafts, which can be called Reader-Based prose. Writer- and Reader-Based prose have been distinguished by Flower (1981):

> In function, Writer-Based prose is a verbal expression written by a writer to himself and for himself. It is the record and the working of his own verbal thought. In its structure, Writer-Based prose reflects the associative, narrative path of the writer's own confrontation with her subject. In its language, it reveals her use of privately loaded terms and shifting but unexpected contexts for her statements. In contrast, Reader-Based prose is a deliberate attempt to communicate something to a reader. To do that, it creates a shared language and shared context between writer and reader. It also offers the reader an issue-centered rhetorical structure rather than a replay of the writer's discovery process. In its language and structure, Reader-Based prose reflects the purpose of the writer's thought; Writer-Based prose tends to reflect its process. Good writing, therefore, is often a cognitively demanding transformation of the natural but private expressions of Writer-Based thought into a structure and style adapted to a reader. (p. 269)

Third, since the subjects wrote their compositions under test conditions, they probably did not have sufficient time to reflect on the assignment and to plan the content and organization. Consequently, they may have experienced a cognitive overload due to the fact that they were attending to the many factors simultaneously. Bereiter and Scardamalia (1983) have discussed this phenomenon observed in children first learning to write. They point out that writers frequently encounter a capacity limit problem and become engrossed with the mechanical aspects of the composing process, leaving little or no time for concerns with content. Flower and Hayes (1980) offer a concise description of the factors that can cause a cognitive overload during the composing process:

> A writer in the act of discovery is hard at work searching memory, forming concepts, and forging a new structure of ideas, while at the same time trying to juggle all the constraints imposed by his or her purpose, audience, and language itself. Discovery, the event, and its product and new insights are only the end result of a complicated intellectual process. (p. 21)

A Final Comment

Thus far, nothing has been written concerning criterion-referenced information on these subjects' performance. According to Matalene, a score of 120 on the RET should be passing. In terms of South Carolina's departmental standards only one of these subjects received a score of 120; the mean RET score was 103.69. These L2 subjects are in their first university-level composition course after having attained

requisite proficiency levels to satisfy their intensive English requirements. The average PROFILE Holistic score for the subject pool was 68.90, which falls into the fair writing ability proficiency stratum defined by Jacobs et al. (1981):

> Proficiency Evaluation: Will probably experience great difficulty completing writing requirements in subject matter courses. May be unable to compete fairly with native writers of English. Look at Profile to identify areas of strengths and weaknesses.
>
> Undergraduate Placement Recommendation: Should have at least one preparatory course in composition before taking college-level English courses or subject matter courses that require much writing.
>
> Graduate Placement Recommendation: Should have at least one basic composition course before entering graduate courses with writing requirements. May also need course in research/technical writing.

Although the purpose of this research was not to equate the RET and the Profile scoring guide, the two instruments seem to produce a concordance of criterion-related information in terms of assessment, placement, exemption, certification, and promotion. The RET results indicate that this subject pool is not ready for main-stream composition. That is not to say that they will in fact be channeled subsequently into main-stream composition, although a certain number will desire to enter composition courses for native speakers of English and will eventually demonstrate that capability. It is also true that

many smaller institutions lack parallel ESL sequences and foreign students must enter the "normal" composition courses. In our institution, if these subjects pass their first non-remedial course, Basic Composition for Foreign Students, they can then take in sequence a typical and common array of composition courses for non-native students: Expository Writing, Technical Writing, and Advanced Composition for Foreign Students.

TABLE 1

SUMMARY STATISTICS

VARIABLE	NO. OF SUBJECTS	NO. OF POINTS (Items)	MEAN	S.D.	TYPE OF RELIABILITY	COEFFICIENT OF RELIABILITY	100-ITEM RELIABILITY
TAS	45	45	32.67	6.16	KR-20	.78	.89
RET	45	150 (30x5)	103.69	9.07	Split-half	.56	.81
HOLISTIC	45	100	68.90	9.23	Interrater	.70	

TABLE 2

VALIDITY COEFFICIENTS

	Maximum Validity Possible $\sqrt{r_{tt(x)}}$	Internal Construct Validity Proportion	r_{xy} With Holistic Criterion
TAS	.88	.67	.44
RET	.75	.37	.40

TABLE 3

STEPWISE MULTIPLE REGRESSION

DEPENDENT VARIABLE-HOLISTIC SCORE (TWO GRADER/AVG)

VARIABLE ENTERED	R SQUARE	R SQUARE IMPROVEMENT	F RATIO$_{Tot}$	DF	PROB
TAS	.19	.19	10.60	1,43	.0022
RET	.34	.15	10.82	2,44	.0022

TABLE 4

STEPWISE MULTIPLE REGRESSION

DEPENDENT VARIABLE-HOLISTIC SCORE (GRADER ONE)

VARIABLE ENTERED	R SQUARE	R SQUARE IMPROVEMENT	F RATIO$_{Tot}$	DF	PROB
TAS	.25	.25	14.77	1,43	.0004
RET	.36	.11	11.91	2,44	.0001

TABLE 5

STEPWISE MULTIPLE REGRESSION

DEPENDENT VARIABLE-HOLISTIC SCORE (GRADER TWO)

VARIABLE ENTERED	R SQUARE	R SQUARE IMPROVEMENT	F RATIO$_{Tot}$	DF	PROB
RET	.13	.13	6.47	1,43	.0146
TAS	.24	.11	6.71	2,44	.0030

TABLE 6

CORRELATION MATRIX

	1	2	3	4	5	6	7	8	9	10
1. TAS	1.00									
2. RET Revision	.13	1.00								
3. RET Editing	-.18	-.06	1.00							
4. RET Total	-.01	.74	61	1.00						
5. PROFILE Content	.24	.32	.05	.28	1.00					
6. PROFILE Organization	.33	.43	.12	.42	.81	1.00				
7. PROFILE Vocabulary	.46	.27	.14	.31	.68	.68	1.00			
8. PROFILE Language Use	.51	.23	.11	.26	.62	.66	.84	1.00		
9. PROFILE Mechanics	.35	-.00	.21	.14	.50	.53	.77	.70	1.00	
10. PROFILE Total (Holistic Score)	.44	.33	.15	.36	.84	.85	.89	.89	.75	1.00

$r = .26$ df 43
$p < .05$, two-tailed test

TABLE 7

STEPWISE MULTIPLE REGRESSION

DEPENDENT VARIABLE HOLISTIC SCORE

VARIABLE ENTERED	R SQUARE	R SQUARE IMPROVEMENT	F RATIO$_{Tot}$	DF	PROB
Vocabulary	.80	.80	179.27	1,43	.0001
Organization	.91	.11	227.09	2,42	.0001
Syntax	.94	.03	240.53	3.41	.0C01
Organization Replaced by Content	.95	.01	266.54	3.41	.0001
Organization	.96	.01	283.02	4,40	.0001
Mechanics	.97	.01	259.06	5,39	.0001

REFERENCES

Bereiter, C., and M. Scardamalia. 1983. Does learning to write have to be so difficult? In A. Freedman, I. Pringle, and J. Yalden (Eds.), Learning to write: first language/second language. London: Longman Group Limited.

Braddock, R., R. Lloyd-Jones, and L. Schoer. 1963. Research in written composition. Urbana, IL: National Council of Teachers of English.

Breland, H., and J. Gaynor. 1979. A comparison of direct and indirect assessments of writing skill. Journal of Educational Measurement, 16:119-128.

Davidson, D. 1978. Teacher's manual for test of ability to subordinate. New York: Language Innovations, Inc.

Davidson, D. 1978. Test of ability to subordinate. New York: Language Innovations, Inc.

Davis, B., M. Scriven, and S. Thomas. 1981. The evaluation of composition instruction. Pt. Reyes, CA: Edgepress.

Diederich, P. 1974. Measuring growth in English. Urbana, IL: National Council of Teachers of English.

Flower, L. 1981. Writer-based prose: a cognitive basis for problems in writing. In G. Tate and E. P. J. Corbett (Eds.), The writing teacher's sourcebook. New York: Oxford University Press.

Flower, L., and J. Hayes. 1980. The cognition of discovery: defining a rhetorical problem. College Composition and Communication, 31: 21-32.

Henning, G. Language test development. Rowley, MA: Newbury House Publishers. In press.

Hogan, T., and C. Mishler. 1980. Relationships between essay tests and objective tests of language skills for elementary school students. Journal of Educational Measurement, 17:219-227.

Jacobs, H., et al. 1981. Testing ESL composition: a practical approach. Rowley, MA: Newbury House Publishers.

Matalene, C. 1982. Objective testing: politics, problems, possibilities. College English, 44:368-381.

Michael, W., and P. Shaeffer. 1979. A comparison of the validity of the Test of Standard Written English (TSWE) and the California State University and Colleges English Placement Test (CSUC-EPT) in the prediction of grades in a basic English composition course and of overall freshman-year grade point average. Educational and Psychological Measurement, 39:131-145.

Moss, P., N. Cole, and C. Khampalikit. 1982. A comparison of procedures to assess written language skills in grades 4, 7, and 10. Journal of Educational Measurement, 19:37-48.

Stiggins, R. 1982. An analysis of direct and indirect methods of writing assessment. Research in the Teaching of English, 16:57-69.